AX

PCT WORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification 6 : C12N 15/31, 15/00, 15/11, C07K 14/315, C07H 21/04, A61K 39/40, 39/395, G01N 33/569, G11B 5/00

A1

(11) International Publication Number: WO 98/26072

(43) International Publication Date: 18 June 1998 (18.06.98)

(21) International Application Number: PCT/US97/22578

(22) International Filing Date: 9 December 1997 (09.12.97)

(30) Priority Data: 60/036,281 13 December 1996 (13.12.96) US

(71) Applicant (*for all designated States except US*): ELI LILLY AND COMPANY [US/US]; Lilly Corporate Center, Indianapolis, IN 46285 (US).

(72) Inventors; and
(75) Inventors/Applicants (*for US only*): BALTZ, Richard, H. [US/US]; 6438 North Olney, Indianapolis, IN 46220 (US). BURGETT, Stanley, G. [US/US]; 818 Bitter Bark Lane, Indianapolis, IN 46227 (US). DEHOFF, Bradley, S. [US/US]; 7727 Maradona Drive South, Indianapolis, IN 46214 (US). HOSKINS, Jo, Ann [US/US]; 8229 Tern Court, Indianapolis, IN 46256 (US). JASKUNAS, Stanley, R., Jr. [US/US]; 16 Algonquian Drive, Natick, MA 07160 (US). MILLS, Bradley, J. [US/US]; 321 South Poplar, Morristown, IN 46161 (US). NORRIS, Franklin, H. [US/US]; 5518 South Arlington Avenue, Indianapolis, IN 46237 (US). PEERY, Robert, B. [US/US]; 372 Sycamore Street, Brownsburg, IN 46112 (US). ROCKEY, Pamela, K. [US/US]; 679 Canary Creek Drive, Franklin, IN 46131 (US). ROSTECK, Paul, R., Jr. [US/US]; 4247 Saffron Drive, Indianapolis, IN 46237 (US). SKATRUD, Paul, L. [US/US]; 5233 Tony Drive, Greenwood, IN 46143 (US). SMITH, Michele, C. [US/US]; 4515 North Lakeridge Drive, Indianapolis, IN 46234 (US). SOLENBERG, Patricia, J. [US/US]; 10116 Seabreeze Way, Indianapolis, IN 46256 (US). TREADWAY, Patti, J. [US/US]; 4268 West Fox Ridge Avenue, Greenwood, IN 46143 (US). YOUNG BELLIDO, Michele, L. [CA/US]; 4925 Common Vista Court, Indianapolis, IN 46220 (US).

(74) Agents: WEBSTER, Thomas, D. et al.; Eli Lilly and Company, Lilly Corporate Center, Indianapolis, IN 46285 (US).

(81) Designated States: AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, CA, CH, CN, CU, CZ, DE, DK, EE, ES, FI, GB, GE, GH, GM, HU, ID, IL, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MD, MG, MK, MN, MW, MX, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TR, TT, UA, UG, US, UZ, VN, YU, ZW, ARIPO patent (GH, GM, KE, LS, MW, SD, SZ, UG, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, ML, MR, NE, SN, TD, TG).

Published
With international search report.
Before the expiration of the time limit for amending the claims and to be republished in the event of the receipt of amendments.

(54) Title: STREPTOCOCCUS PNEUMONIAE DNA SEQUENCES

(57) Abstract

Nucleotide fragments of the genome of the bacterium *Streptococcus pneumoniae* are provided. Also provided are ORFs encoded by said genome.

FOR THE PURPOSES OF INFORMATION ONLY

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AL	Albania	ES	Spain	LS	Lesotho	SI	Slovenia
AM	Armenia	FI	Finland	LT	Lithuania	SK	Slovakia
AT	Austria	FR	France	LU	Luxembourg	SN	Senegal
AU	Australia	GA	Gabon	LV	Latvia	SZ	Swaziland
AZ	Azerbaijan	GB	United Kingdom	MC	Monaco	TD	Chad
BA	Bosnia and Herzegovina	GE	Georgia	MD	Republic of Moldova	TG	Togo
BB	Barbados	GH	Ghana	MG	Madagascar	TJ	Tajikistan
BE	Belgium	GN	Guinea	MK	The former Yugoslav Republic of Macedonia	TM	Turkmenistan
BF	Burkina Faso	GR	Greece			TR	Turkey
BG	Bulgaria	HU	Hungary	ML	Mali	TT	Trinidad and Tobago
BJ	Benin	IE	Ireland	MN	Mongolia	UA	Ukraine
BR	Brazil	IL	Israel	MR	Mauritania	UG	Uganda
BY	Belarus	IS	Iceland	MW	Malawi	US	United States of America
CA	Canada	IT	Italy	MX	Mexico	UZ	Uzbekistan
CF	Central African Republic	JP	Japan	NE	Niger	VN	Viet Nam
CG	Congo	KE	Kenya	NL	Netherlands	YU	Yugoslavia
CH	Switzerland	KG	Kyrgyzstan	NO	Norway	ZW	Zimbabwe
CI	Côte d'Ivoire	KP	Democratic People's Republic of Korea	NZ	New Zealand		
CM	Cameroon			PL	Poland		
CN	China	KR	Republic of Korea	PT	Portugal		
CU	Cuba	KZ	Kazakstan	RO	Romania		
CZ	Czech Republic	LC	Saint Lucia	RU	Russian Federation		
DE	Germany	LI	Liechtenstein	SD	Sudan		
DK	Denmark	LK	Sri Lanka	SE	Sweden		
EE	Estonia	LR	Liberia	SG	Singapore		

STREPTOCOCCUS PNEUMONIAE DNA SEQUENCES

This invention provides DNA sequences from the *Streptococcus pneumoniae* genome, and methods of use of DNA fragments originating therefrom in a variety of biological and pharmaceutical applications.

The recent emergence of widespread antibiotic resistance in common pathogenic bacterial species has justifiably alarmed the medical and research communities. Frequently these organisms are co-resistant to several different antibacterial agents. Particularly problematic has been the emergence and rapid spread of penicillin resistance in *Streptococcus pneumoniae,* which frequently causes upper respiratory tract infections. Resistance to penicillin in this organism can be due to modifications of one or more of the penicillin-binding proteins (PBPs). Combating the phenomenon of increasing resistance to antibiotic agents among pathogenic organisms such as *Streptococcus pneumoniae* will require intensified research into the fundamental molecular biology of such organisms. Greater knowledge about the molecular biology of pathogenic organisms will lead to new antibacterial agents having novel and effective actions.

While inroads in the development of new antibiotics and new targets for antibiotic compounds have been made with a variety of microorganisms, progress has been less apparent in *Streptococcus pneumoniae*. In part, *Streptococcus pneumoniae* presents a special case because this organism is highly recombinogenic and readily takes up exogenous DNA from its surroundings. Thus, the need for new antibacterial compounds and new targets for antibacterial therapy in *Streptococcus pneumoniae* is more acute than in other organisms.

The present invention relates to the genome of *S. pneumoniae*. The genomic information disclosed by the present invention enables: (1) preparation of molecular hybridization probes for use in PCR amplification of genes and regulatory regions, physical mapping, sequencing, mutagenesis, and mutation analysis, (2) homology comparisons with the genomes and open reading frames (ORFs) of other organisms, (3) creation of specifically mutated strains of *S. pneumoniae* wherein the mutation is targeted to any site or sites in the DNA sequence disclosed herein, (4) identification of *S. pneumoniae* promoters and other gene regulatory sequences, (5) identification of proteins/ORFs encoded by *S. pneumoniae*, (6) identification of virulence genes in *S. pneumoniae*, (7) determination of the biological function of proteins/ORFs and RNAs encoded by *S. pneumoniae*, (8) production of kits useful for determining gene function in the cell, and kits for isolating and analyzing genes that are mutated in antibiotic resistant clinical isolates of *S. pneumoniae*, (9) production of proteins and RNAs encoded by *S. pneumoniae*, (10) production of antibodies against proteins and other antigens encoded by *S. pneumoniae*, (11) methods to identify compounds that bind to proteins and RNAs encoded by *S. pneumoniae* as potential new antibiotic compounds.

In another embodiment the invention relates to substantially purified proteins encoded by the *S. pneumoniae* genome.

Table 1 summarizes the proteins and nucleic acids disclosed herein, contigs, SEQ ID NO's and predicted functions.

"Genome" refers to the full complement of chromosomal and extra-chromosomal DNA within a cell. The genome comprises the genetic blueprint for all proteins and RNAs encoded by the cell or organism.

"ORF" (i.e. "open reading frame") designates a region of genomic DNA beginning with a Met or other initiation codon and terminating with a translation stop codon, potentially encoding a protein product. "Partial ORF" means a portion of an ORF as disclosed herein such that the initiation codon, the stop codon, or both are not disclosed.

"DNA chip" or "Bio Chip" or "Bio DNA chip" refers to a solid matrix or support onto which is applied an array of oligonucleotides, or nucleotide sequences, or gene fragments, or genomic fragments, of *S. pneumoniae* which may further comprise a layer of *S. pneumoniae* cells suspended thereover in a semisolid medium such as agar or agarose.

"Consensus sequence" refers to an amino acid or nucleotide sequence that may suggest the biological function of a protein, DNA, or RNA molecule. Consensus sequences are identified by comparing proteins, RNAs, and gene homologs from different species.

"Contiguous fragment building" or "Contiguous fragment" or "Contig" refers to the process and result, respectively, by which a fragment of DNA is assembled from smaller constituent DNA fragments by arranging the constituent pieces in their correct order and register such that the resulting contiguous fragment accurately depicts the native DNA sequence from which the smaller fragments originated.

"Computer readable medium" includes, for example, a floppy disc, hard disc, random access memory, read only memory, and CD-ROM.

The terms "cleavage" or "restriction" of DNA refers to the catalytic cleavage of the DNA with a restriction enzyme that acts only at certain sequences in

the DNA (viz. sequence-specific endonucleases). The various restriction enzymes used herein are commercially available and their reaction conditions, cofactors, and other requirements are used in the manner well known to one of ordinary skill in the art. Appropriate buffers and substrate amounts for particular restriction enzymes are specified by the manufacturer or can be found in the literature.

"Diagnostics" as used herein relates to *in vitro* or *in vivo* diagnosis for disease states or biological status in mammals, preferably humans.

"Therapeutics" and "therapeutic/diagnostic combinations" means the treatment, or diagnosis and treatment, of disease states or biological status by *in vivo* administration to mammals, preferably humans, of compositions of the present invention, for example, antibodies.

"Essential genes" or "essential ORFs" or "essential proteins" refer to genomic information or the protein(s) or RNAs encoded therefrom, which, when disrupted by knockout mutation, or by other mutation, produce inviability in cells harboring said mutation.

"Non-essential genes" or "non-essential ORFs" or "non-essential proteins" refer to genomic information or the protein(s) or RNAs encoded therefrom, which, when disrupted by knockout mutation, or other mutation, do not result in inviability of cells harboring said mutation.

"Minimal gene set" refers to a genus of about 256 genes that are conserved among different bacteria such as *M. genitalium* and *H. influenzae*. The minimal gene set appears to be necessary and sufficient to sustain life. *See e.g.* A. Mushegian and E. Koonin, "A minimal gene set for cellular life derived by comparison of complete bacterial genomes" *Proc. Nat. Acad. Sci.* 93, 10268 - 273 (1996).

The term "fragment thereof" denotes a fragment of a nucleic acid molecule described herein, wherein said fragment comprises a region of contiguity within said nucleic acid of at least 15 base pairs. The term may also refer to a peptide of at least 5 contiguous amino acid residues of a protein disclosed herein.

The term "plasmid" refers to an extrachromosomal genetic element. The starting plasmids herein are either commercially available, publicly available on an unrestricted basis, or can be constructed from available plasmids in accordance with published procedures. In addition, equivalent plasmids to those described are known in the art and will be apparent to the ordinarily skilled artisan.

"Recombinant DNA cloning vector" as used herein refers to any autonomously replicating agent, including, but not limited to, plasmids and phages, comprising a DNA molecule to which one or more additional DNA segments can or have been added.

The term "recombinant DNA expression vector" as used herein refers to any recombinant DNA cloning vector, for example a plasmid or phage, in which a promoter and other regulatory elements are present to enable transcription of the inserted DNA.

The term "vector" as used herein refers to a nucleic acid compound used for introducing exogenous DNA into host cells. A vector comprises a nucleotide sequence which may encode one or more protein molecules. Plasmids, cosmids, viruses, and bacteriophages, in the natural state or which have undergone recombinant engineering, are examples of commonly used vectors.

The terms "complementary" or "complementarity" as used herein refers to the capacity of purine and pyrimidine nucleotides to associate through hydrogen bonding in double

stranded nucleic acid molecules. The following base pairs are complementary: guanine and cytosine; adenine and thymine; and adenine and uracil.

"Oligonucleotide" refers to a short polymeric nucleotide chain comprising from about 2 to 25 nucleotides.

"Isolated nucleic acid compound" refers to any RNA or DNA sequence, however constructed or synthesized, which is locationally distinct from its natural location.

A "primer" is a nucleic acid fragment which functions as an initiating substrate for enzymatic or synthetic elongation of a nucleic acid molecule.

The term "promoter" refers to a DNA sequence which directs transcription of DNA to RNA.

A "probe" as used herein is a labeled nucleic acid compound which can be used to hybridize with another nucleic acid compound.

The term "hybridization" or "hybridize" as used herein refers to the process by which a single-stranded nucleic acid molecule joins with a complementary strand through nucleotide base pairing.

"Recorded" as used herein refers to a process for storing information on a computer readable medium.

"Substantially identical" means a sequence having sufficient homology to hybridize under high stringency conditions and/or at least 90% identity at the nucleotide or amino acid sequence level to a sequence disclosed herein.

"Substantially purified" when used in reference to a protein or peptide means that the molecule has been largely, but not necessarily wholly, separated and purified from other cellular and non-cellular components. Typically a protein is substantially pure when it is at least about 60% by weight, free from other naturally occurring organic molecules. Preferably the purity is at least about 75%, more

preferably at least about 90%, and most preferably at least about 99% by weight pure.

"Selective hybridization" refers to hybridization under conditions of high stringency. Hybridization of nucleic acid molecules depends upon factors such as the degree of complementarity, stringency of hybridization conditions, and the length of hybridizing strands.

The term "stringency" relates to nucleic acid hybridization conditions. High stringency conditions disfavor non-homologous base pairing. Low stringency conditions have the opposite effect. Stringency may be altered, for example, by changes in temperature and salt concentration. Typical high stringency conditions comprise hybridizing at 50°C to 65°C in 5X SSPE and 50% formamide, and washing at 50°C to 65°C in 0.5X SSPE; typical low stringency conditions comprise hybridizing at 35°C to 37°C in 5X SSPE and 40% to 45% formamide and washing at 42°C in 1X-2X SSPE.

"SSPE" denotes a hybridization and wash solution comprising sodium chloride, sodium phosphate, and EDTA, at pH 7.4. A 20X solution of SSPE is made by dissolving 174 g of NaCl, 27.6 g of $NaH_2PO4 \cdot H_2O$, and 7.4 g of EDTA in 800 ml of H_2O. The pH is adjusted with NaOH and the volume brought to 1 liter.

"SSC" denotes a hybridization and wash solution comprising sodium chloride and sodium citrate at pH 7. A 20X solution of SSC is made by dissolving 175 g of NaCl and 88 g of sodium citrate in 800 ml of H_2O. The volume is brought to 1 liter after adjusting the pH with 10N NaOH.

"Virulence gene" as used herein means a gene from a pathogenic organism such as *S. pneumoniae* that is required for infection and/or pathogenicity *in vivo*. Some virulence

genes are induced during infection of a host; others are expressed exclusively during *in vivo* infection.

The *Streptococcus pneumoniae* genome contains about 2.2 million nucleotide base pairs and comprises about 2000 to 3000 ORFs and other genes. This invention provides, among other things, contiguous fragments, genes, and proteins from the *S. pneumoniae* genome (SEQ ID NO:1 through SEQ ID NO:228).

Strain differences in *S. pneumoniae* may be associated with nucleotide sequence differences in one or more of the genomic fragments disclosed herein. Sequences that are substantially identical to the sequences disclosed herein are intended to be within the scope of the invention.

The sequence fragments disclosed herein provide a wide variety of utilities. For example, the fragments may be used to identify regions of the *S. pneumoniae* genome that are expressed as proteins (viz. transcribed into mRNA). The genomic fragments disclosed herein can also be used to examine differential expression of *S. pneumoniae* genes under diverse environmental conditions, as occurs, for example, with the expression of virulence genes during in vivo infection of a host organism. Also contemplated by the invention are: (1) preparation of molecular hybridization probes for use in physical mapping, sequencing, mutagenesis, mutation analysis, (2) homology comparisons of the sequences disclosed herein with the genomes and ORFs of other organisms, (3) creation of specifically mutated strains of *S. pneumoniae* wherein the mutation is targeted to any site in the DNA sequence disclosed herein, (4) identification of *S. pneumoniae* promoters and other gene regulatory sequences, (5) identification of proteins and RNAs encoded by *S. pneumoniae*, (6) amplification of *S. pneumoniae* genes using the PCR, and (7) production of kits for isolating and

analyzing genes that are mutated in antibiotic resistant clinical isolates of *S. pneumoniae*.

Genome Analysis

In one embodiment, the invention comprises the ORFs and fragments thereof encoded by the nucleotide sequences disclosed herein. Some of the nucleotide sequences disclosed herein encode ORFs and fragments of ORFs (Table 1). The ORFs or fragments thereof were identified by translation of the nucleic acid sequences disclosed herein. The biological function of a protein disclosed in Table 1 was determined by homology comparison with known proteins from other organisms. A number of computer programs are available to assist in homology comparisons, for example Genemark (Borodovsky and McIninch, *Computers Chem.* 17(2), 123, 1993).

Computer-Related Applications

The nucleotide and/or amino acid sequence information of this invention may be provided in a variety of media to facilitate use. In one embodiment the present invention comprises one or more of the sequences disclosed herein recorded on a computer readable medium. A variety of media are contemplated, for example, magnetic storage media such as floppy discs, hard disc storage, magnetic tape, and CD-ROM. A skilled artisan can readily adopt any presently known method for recording information on a computer readable medium to generate manufactures comprising the nucleotide or amino acid sequence information of the present invention. These embodiments are contemplated within the scope of this invention.

The choice of a data storage structure will generally be based on the means chosen to access the stored information. A variety of data processor programs and formats can be used to store the sequence information of the

invention on computer readable medium. For example, the sequence can be represented in a word processing text file that is formatted in commercially available software such as WordPerfect and MicroSoft Word, or it can be represented in the form of a text only file such as ASCII.

Having *S. pneumoniae* genomic sequence information in a computer readable format enables a skilled artisan to access the information for a variety of purposes. For example, computer-assisted searching algorithms may be used to identify open reading frames, and ascertain biological function based on homology to known proteins from other organisms. Suitable algorithms for sequence comparisons include BLAST (Altschul *et al.*, *J. Mol. Biol.* 215, 403-410, 1990) and BLAZE (Brutlag *et al.*, *Comp. Chem.* 17, 203-207 (1993). For identification of ORFs a number of commercially available software programs are suitable, such as FRAMES (Genetic Center Group, Madison, WI).

The genomic information of this invention in computer-readable form can be manipulated further using bioinformatics to identify the biological function of proteins encoded by ORFs as well as the cellular location of said proteins. The skilled artisan will recognize several computer-assisted algorithms for this purpose, for example, PSORT which is useful for determining the likely location of a protein within a cell (*See* K. Nakai & M. Kanehisa. "Expert system for predicting protein localization sites in Gram-negative bacteria", *Proteins: Structure, Function, and Genetics*, 11, 95-110 (1991).

Open Reading Frames and Proteins

The invention also provides proteins encoded by the *S. pneumoniae* genome in substantially purified form (*See* Table 1). The proteins are classified herein as (1) Hypothetical,

(2) Cell wall biosynthetic, (3) External target, or (4) Minimal gene set proteins.

Cells that carry knockout mutations in proteins of the hypothetical class are nonviable. Loss of viability suggests that these proteins may be essential for viability. Two such proteins, whose genes map to contigs m014 and m016, correspond respectively to *Haemophillus influenzae* ORFs HI1146 and HI1648. Two other hypothetical proteins, yyaF and ywbL, correspond to a GTP binding protein and transcriptional regulator, respectively.

The proteins of this invention can be used to raise antibodies. Antibodies against the hypothetical class of proteins are especially attractive. In targeting presumptively essential cellular functions, antibodies against "hypothetical proteins" could have therapeutic or prophylactic applications. Additionally, the "hypothetical" proteins can be used to screen for agents that bind or otherwise interact with said proteins. Such agents could lead to the identification of new antibacterial agents.

Proteins classified in Table 1 as cell wall biosynthetic proteins, and external target proteins, were identified by homology with known proteins. These proteins are useful for identifying agents that bind and inhibit bacterial growth. Therefore, in another embodiment of the invention, the proteins of these classifications are prepared, preferably by recombinant means as described herein, substantially purified, and used in a screen to identify compounds that bind and/or inhibit the activity of said proteins. A variety of suitable screens are contemplated for this purpose. For example, the protein(s) can be labeled by known techniques such as radiolabeling or fluorescent tagging, or by labeling with biotin/avidin; thereafter binding of a test compound to a labeled protein

can be determined by any suitable means, well known to the skilled artisan.

The proteins categorized as "minimal gene set" are homologous to a set of highly conserved proteins found in other bacteria. The minimal gene set proteins are thought to be essential for viability, and are useful targets for the development of new antibacterial compounds.

DNA Chips and Applications

The nucleic acids disclosed herein, or subfragments thereof, may be arrayed on any suitable solid surface, thereby constructing a "chip." DNA chip hybridizations provide greater sensitivity than do conventional hybridization means, such as Southern hybridization or Northern hybridization. DNA chips are useful for a variety of purposes, for example, in mutation and gene expression analysis, and in probing the structure, function, and expression of the genome. This aspect of the invention relates to any one or more of the DNA fragments disclosed herein, wherein said fragments are attached to a solid support (i.e."chip" or "DNA chip" or "Bio chip"). Attachment of a nucleic acid to a support can be, but is not necessarily, accomplished by chemical or enzymatic means.

In one embodiment, DNA fragments of this invention are arrayed onto a solid support as a means for assessing gene expression in *S. pneumoniae*. The DNA fragments attached to a chip may be of any size that is suitable for hybridization to other nucleic acid molecules such as cDNAs, genomic DNAs, or RNAs. Suitably-sized DNA fragments are from 10 nucleotide residues to approximately several thousand residues. The preferred length is about 50 to 500 nucleotides.

Analysis of gene expression using the chips of this invention is assessed by hybridization of a chip to RNA samples, or cDNA samples prepared from *S. pneumoniae* grown

under any suitable conditions. Preferred samples for hybridization to a chip comprise cDNA. Methods for preparing RNA or cDNA are well known in the art.

A variety of suitable methods are known for fixing DNA fragments to solid support matrices [*See e.g.* D. Stimpson *et al.* "Real-time detection of DNA hybridization and melting on oligonucleotide arrays by using optical wave guides" *Proc. Nat. Acad. Sci.* 92, 6379 (1995)] Preferred surfaces for producing a chip are glass or polystyrene. Convenient surfaces are microscope slides, or cover slips (Corning), treated with silicon or silane to minimize non-specific binding by DNA or proteins. Also suitable for this purpose are 96-well microtiter plates.

A light-directed method may be used for attaching oligonucleotides, enabling nucleotide synthesis directly on the solid surface using photolabile 5'protected N-acyl-deoxynucleotide phosphoramidites and surface linker chemistry (*See* Pease *et al.* "Light-generated oligonucleotide arrays for rapid DNA sequence analysis" *Proc. Nat. Acad. Sci.* 91, 5022-5026, 1994). Alternatively, DNA fragments can be bound to a surface via interaction with a specific DNA binding protein. Any suitable DNA binding protein may be used, for example bacteriophage DNA binding proteins, Adenovirus binding protein, the *E. coli* lac-repressor protein, or l-repressor protein. DNA binding proteins are attached to the surface of a chip by covalent chemical binding, essentially as described in U.S. Patent 5,561,071, the entire contents of which is incorporated by reference. The latter method requires that DNA fragments contain a recognition sequence that enables binding by the DNA binding protein. Specific sequences for a number of DNA binding proteins are known. Methods for incorporating specific binding sequences into the genomic DNA fragments disclosed herein are well known in the cloning arts.

DNA chip technology enables monitoring *S. pneumoniae* gene expression on a genome-wide level. This feature of the invention is particularly attractive for identifying (1) genes that are expressed or not expressed during the life cycle or infection cycle of *S. pneumoniae*, and (2) changes in gene expression that correlate with environmental change.

For example, virulence genes in *S. pneumoniae* can be identified by the DNA chip method disclosed herein. Identification of virulence genes in *S. pneumoniae* will provide new targets for developing novel antibiotics. For this aspect of the invention any suitable encapsulated strain of *S. pneumoniae* is introduced into a mouse, for example, by intraperitoneal injection, or by introduction directly into the lungs, or by any other suitable method. Approximatly 2 days after infection a peripheral blood titre level is reached of about 10^8 *S. pneumoniae* cells/ml. Cells recovered from peripheral blood, or other suitable tissue, are used in identifying virulence genes. For this purpose, cDNAs are prepared from cells recovered from an *in vivo* infection and from cells grown *in vitro*. After labeling, the cDNAs are hybridized against the DNA chip(s) disclosed herein. Genomic fragments that hybridize to the *in vivo* probe but not to the *in vitro* probe identify candidate virulence genes.

Also contemplated by this aspect of the invention is a method for analyzing gene expression in *S. pneumoniae* cells grown or harvested from any desireable *in vitro* or *in vivo* environment, wherein said environment may include compounds whose effects on gene expression are to be determined.

In another embodiment, the present invention relates to a DNA bio-chip, useful for correlating DNA sequence with biological function. The bio-chip comprises an array of the genomic DNA fragments disclosed herein, or portions thereof, attached to the surface of any suitable solid support

material. The bio chip further comprises a layer of competent *S. pneumoniae* cells suspended over the DNA array in any suitable semi-solid medium such as agar or agarose. The cells suspended on the bio chip comprise known or unknown mutant strains, or they may be wild-type cells. The cell layer is in contact with the DNA matrix such that DNA on the chip can be taken up by the cells.

The bio-chip is useful for several purposes. For example, the bio-chip can be used to localize an unknown mutation to a specific region of the genome by complementation. The bio-chip enables correlating a phenotype with a genetic locus. For example, mutant cells harboring one or more mutations and having at least one screenable or selectable phenotype can be applied to a bio chip and subjected to an environment that allows for selection, or for screening by complementation. If said phenotype is the result of a chromosomal mutation or mutations that map to a genomic fragment present on the chip, DNA uptake by the cells and repair of the mutation by recombination will be identifiable by a suitable screen or selection.

In a preferred embodiment, the bio-chip is overlayed with competent *S. pneumoniae* cells. Methods for preparing competent cells are known (*See e.g.* LeBlanc *et.al. Plasmid* 28, 130-145, 1992; Pozzi *et al. J. Bacteriol.*178, 6087-6090, 1996).

Other embodiments of this aspect of the invention are contemplated. For example the genomic fragments disclosed herein could be prepared and dispensed into individual wells of a 96-well micro titre plate. Competent *S. pneumoniae* cells could then be added to the wells under conditions suitable for DNA uptake followed by plating onto any suitable selection or screening medium, for example an agar

plate containing suitable growth and/or selection/screening components.

Diagnostic Kits and Assays

The present invention further relates to kits and assays that can be used for rapid and efficient detection of *S. pneumoniae* cells. Also contemplated are kits for detecting mutations carried by *S. pneumoniae* cells. Kits of this nature are particularly attractive in the clinical environment where knowledge about the identity of a pathogen and/or of the basis for resistance to antibiotic treatments is essential for effective medical treatment. In the long term, knowledge of the mutations that lead to resistance will enable the design of new antibacterial agents.

A kit for detecting *S. pneumoniae* cells can be based on antibody recognition of *S. pneumoniae* specific antigens or epitopes, or by nucleic acid hybridization techniques for the detection of *S. pneumoniae* specific nucleic acid molecules.

A variety of embodiments are contemplated in this aspect of the invention. In one embodiment a kit is provided for detecting mutations in drug-resistant *S. pneumoniae*. For this purpose, DNA is prepared from a resistant isolate and from a wild-type strain. In a preferred embodiment, the polymerase chain reaction (i.e. PCR) is used to amplify DNA samples representing any one or all of the genomic fragments disclosed herein. The amplified DNAs from the mutant and wild-type cells are hybridized to a DNA chip having fixed thereon any one or more of the genomic fragments disclosed herein. Amplified DNA samples from the mutant and wild-type strain are labeled by any suitable means, for example using radioisotopes or fluorescent labeling. Hybridization of the amplified DNAs to the chip under conditions that can discriminate single or multiple base pair mismatches enables

the detection of differences between the mutant and wild-type samples. This method identifies a specific fragment of the genome that is altered in the mutant strain. The specific mutation can be determined by conventional DNA sequence analysis.

This aspect of the invention also relates to the detection of *S. pneumoniae* proteins in a sample using antibody molecules raised against any suitable ORF disclosed herein. Antibody detection methods are well known to those skilled in the art including, for example, a variety of radioimmunological assays. (*See e.g.* P. Tijssen, Practice and Theory of Enzyme Immunoassays: Laboratory Techniques in Biochemistry and Molecular Biology, Elsevier Science Publishers, Amsterdam, The Netherlands, 1985).

Test samples suitable for use in this aspect of the invention include but are not limited to biological fluids such as sputum, blood, serum, plasma, urine, and to biopsy samples.

Skilled artisans will recognize that the disclosed method and reagents can be readily incorporated into a kit. For example, a kit would contain one or more receptacles comprising one or more of the following: PCR reagents, DNA chip reagents, labeling reagents, assorted buffers, and/or antibodies.

Production of Antibodies

The proteins of this invention and fragments thereof may be used in the production of antibodies. The term "antibody" as used herein describes antibodies, fragments of antibodies (such as, but not limited, to Fab, Fab', Fab_2', and Fv fragments), and chimeric, humanized, veneered, resurfaced, or CDR-grafted antibodies capable of binding antigens of a similar nature as the parent antibody

molecule from which they are derived. The instant invention also encompasses single chain polypeptide binding molecules.

The production of antibodies, both monoclonal and polyclonal, in animals is well known in the art. *See, e.g.,* C. Milstein, Handbook of Experimental Immunology, (Blackwell Scientific Pub., 1986); J. Goding, Monoclonal Antibodies: Principles and Practice, (Academic Press, 1983). For the production of monoclonal antibodies the process begins with injecting a mouse, or other suitable animal, with an immunogen. The mouse is subsequently sacrificed and cells taken from its spleen are fused with myeloma cells, resulting in a hybridoma that can be cultured *in vitro*. Hybridomas are screened for clones that secrete a single antibody species, specific for the immunogen.

Chimeric antibodies, described in U.S. Patent No. 4,816,567, herein incorporated by reference, teaches methods and vectors for preparing chimeric antibodies. An alternative approach is provided in U.S. Patent No. 4,816,397, the entire contents of which is herein incorporated by reference. This patent teaches co-expression of heavy and light chains in the same host cell.

The method taught in U.S. Patent 4,816,397 has been further refined in European Patent Publication No. 0 239 400. The teachings of this publication are preferred for engineering monoclonal antibodies. In this technology the complementarity determining regions (CDRs) of a human antibody are replaced with the CDRs of a murine monoclonal antibody, thereby converting the specificity of the human antibody to the specificity of the murine antibody.

Single chain antibodies and libraries thereof provide yet another means for genetically engineering antibody molecules. (*See, e.g.* R.E. Bird, *et al., Science* 242:423-426 (1988); PCT Publication Nos. WO 88/01649, WO 90/14430, and WO 91/10737. Single chain antibody technology

involves covalently joining the binding regions of heavy and light chains thereby generating a single polypeptide chain having the binding specificity of an intact antibody molecule.

The antibodies contemplated by the present invention are useful in diagnostics, therapeutics, or in diagnostic/therapeutic combinations.

The proteins of this invention, or suitable fragments thereof, can be used to generate polyclonal or monoclonal antibodies, and various inter-species hybrids, or humanized antibodies, or antibody fragments, or single-chain antibodies. The techniques for producing antibodies are well known to skilled artisans. (*See e.g.* A.M. Campbell, Monoclonal Antibody Technology: Laboratory Techniques in Biochemsitry and Molecular Biology, Elsevier Science Publishers, Amsterdam (1984); Kohler and Milstein, *Nature* 256, 495-497 (1975); Monoclonal Antibodies: Principles & Applications Ed. J.R.Birch & E.S. Lennox, Wiley-Liss, 1995.

A protein or peptide to be used as an immunogen may be administered in an adjuvant by subcutaneous or intraperitoneal injection into, for example, a mouse or a rabbit. For the production of monoclonal antibodies, spleen cells from immunized animals are removed, fused with myeloma cells, such as SP2/0-Ag14 cells, and allowed to become monoclonal antibody producing hybridoma cells in the manner known to the skilled artisan. Hybridomas that secrete the desired antibody molecule can be screened by a variety of well known methods, for example ELISA assay, western blot analysis, or radioimmunoassay (Lutz *et al. Exp. Cell Res.* 175, 109-124 (1988); Monoclonal Antibodies: Principles & Applications Ed. J.R.Birch & E.S. Lennox, Wiley-Liss, 1995).

For some applications it is desireable to have an antibody labeled in some fashion. Procedures for labeling antibody molecules with radioisotopes, affinity labels, such

as biotin or avidin, enzymatic labels, for example horseradish peroxidase, and fluorescent labels such as FITC or rhodamine, are widely known (*See e.g.* Enzyme-Mediated Immunoassay, Ed. T. Ngo, H. Lenhoff, Plenum Press 1985; Principles of Immunology and Immunodiagnostics, R.M. Aloisi, Lea & Febiger, 1988).

Labeled antibodies are useful for a variety of diagnostic applications. In one embodiment, the present invention relates to the use of labeled antibodies to detect the presence of *S. pneumoniae* cells and proteins. Also contemplated are applications that use antibodies, preferably single chain antibodies, directed against a *S. pneumoniae* protein. Proteins identified as "external targets" are preferred for the generation of single chain antibodies. Single chain antibody libraries directed against *S. pneumoniae* surface proteins and cell wall proteins can be produced by applying the phage display technique to crude membrane preparations. Antibodies that recognize and bind to external target proteins and/or cell wall proteins could be used as therapeutic agents to inhibit the growth of *S. pneumoniae*. Alternatively, the antibodies could be used in a screen to identify potential inhibitors of an external target protein. For example, in a competitive displacement assay, an antibody or compound to be tested is labeled by any suitable method. Competitive displacement of an antibody from an antibody-antigen complex by a test compound provides a means to identify new antibacterial compounds.

Protein Production Methods

The present invention relates further to substantially purified proteins encoded by the ORFs disclosed herein (SEQ ID NO:87 through SEQ ID NO:228).

Skilled artisans will recognize that proteins can be synthesized by different methods, for example, chemical

methods or recombinant methods, as described in U.S. Patent 4,617,149, hereby incorporated by reference.

The principles of solid phase chemical synthesis of polypeptides are well known in the art and may be found in general texts relating to this area. *See, e.g.*, H. Dugas and C. Penney, Bioorganic Chemistry (1981) Springer-Verlag, New York, 54-92. Peptides may be synthesized by solid-phase methodology utilizing an Applied Biosystems 430A peptide synthesizer (Applied Biosystems, Foster City, CA) and synthesis cycles supplied by Applied Biosystems. Protected amino acids, such as t-butoxycarbonyl-protected amino acids, and other reagents are commercially available from many chemical supply houses.

The proteins and peptides of the present invention can also be made by recombinant DNA methods. Recombinant methods are preferred if a high yield is desired. Recombinant methods involve expressing a cloned ORF/gene in a suitable host cell. A gene is introduced into a host cell by any suitable means, well known to those skilled in the art. While chromosomal integration of a cloned gene is within the scope of the present invention, it is preferred that a cloned gene be maintained extra-chromosomally, as part of a vector wherein the gene is in operable-linkage to a constitutive or inducible promoter.

Recombinant methods are also useful in overproducing a membrane-bound or membrane-associated protein. In some cases, membranes prepared from recombinant cells that overexpress such proteins provide an enriched source of the protein. Such membranes are useful for evaluating the function of the protein and/or for evaluating inhibitors of the protein.

Expressing Recombinant Proteins in Procaryotic and Eucaryotic Host Cells

Procaryotes are generally used for cloning DNA sequences and for constructing vectors. For example, the *Escherichia coli* K12 strain 294 (ATCC No. 31446) is particularly useful for expression of foreign proteins. Other strains of *E. coli,* bacilli such as *Bacillus subtilis*, enterobacteriaceae such as *Salmonella typhimurium* or *Serratia marcescans*, various Pseudomonas species may also be employed as host cells in cloning and expressing the recombinant proteins of this invention. Also contemplated are various strains of *Streptococcus* and *Streptocmyces*.

For effective expression of a recombinant protein a gene or ORF may be linked to a known promoter sequence. Suitable bacterial promoters include b -lactamase [*e.g.* vector pGX2907, ATCC 39344, contains a replicon and b - lactamase gene], lactose systems [Chang *et al., Nature* (London), 275:615 (1978); Goeddel *et al., Nature* (London), 281:544 (1979)], alkaline phosphatase, and the tryptophan (trp) promoter system [vector pATH1 (ATCC 37695)] designed for the expression of a trpE fusion protein. Hybrid promoters such as the tac promoter (isolatable from plasmid pDR540, ATCC-37282) are also suitable. Promoters for use in bacterial systems also will contain a Shine-Dalgarno sequence operably linked to the DNA encoding the desired polypeptides. These examples are illustrative rather than limiting.

A variety of mammalian cell systems and yeasts are also suitable host cells. The yeast *Saccharomyces cerevisiae* is a commonly used eucaryotic microorganism. Other yeasts such as *Kluyveromyces lactis* are also suitable. For expression of recombinant genes in *Saccharomyces*, the plasmid YRp7 (ATCC-40053), for example, may be used. *See, e.g.*, L. Stinchcomb, *et al., Nature*, 282:39 (1979); J.

Kingsman *et al., Gene,* 7:141 (1979); S. Tschemper *et al., Gene,* 10:157 (1980). Plasmid YRp7 contains the TRP1 gene that provides a selectable marker in a trp1 mutant.

Purification of Recombinantly-Produced Protein

An expression vector carrying an ORF of the present invention is transformed or transfected into a suitable host cell using standard methods. Cells which contain the vector are propagated under conditions suitable for expression of the encoded protein. If the gene is under the control of an inducible promoter then suitable growth conditions would incorporate the appropriate inducer. The recombinantly-produced protein may be purified from cellular extracts of transformed cells by any suitable means.

In a preferred process for protein purification a gene/ORF is modified at the 5' end, or some other position, to incorporate a plurality of histidine residues at the amino terminus of the encoded protein. The "histidine tag" produced thereby enables a single-step protein purification method referred to as "immobilized metal ion affinity chromatography" (IMAC), essentially as described in U.S. Patent 4,569,794, hereby incorporated by reference. The IMAC method enables rapid isolation of substantially pure protein starting from a crude cellular extract.

As skilled artisans will recognize, the proteins of the invention can be encoded by a multitude of different nucleic acid sequences owing to the degeneracy of the genetic code. The present invention further comprises these alternate nucleic acid sequences.

The ribonucleic acid compounds of the present invention may be prepared using the polynucleotide synthetic methods discussed *supra,* or they may be prepared enzymatically using RNA polymerase to transcribe a DNA template.

The most preferred systems for preparing the ribonucleic acids of the present invention employ the RNA polymerase from the bacteriophage T7 or the bacteriophage SP6. These RNA polymerases are highly specific, requiring the insertion of bacteriophage-specific sequences at the 5' end of the template to be transcribed. *See*, J. Sambrook, *et al.*, *supra*, at 18.82-18.84.

This invention also provides nucleic acids, RNA or DNA, which are complementary to the sequences disclosed herein.

The present invention also provides probes and primers useful for a variety of molecular biology techniques including, for example, hybridization screens of genomic or subgenomic libraries, detection and quantification of mRNA species as a means to analyzing gene expression, and amplification of any region of the *Streptococcus pneumoniae* genome disclosed by the sequences herein. A nucleic acid compound is provided comprising any of the sequences disclosed herein, or a complementary sequence thereof, or a fragment thereof, which is at least 15 base pairs in length, and which will hybridize selectively to *Streptococcus pneumoniae* DNA or mRNA. Preferably, the 15 or more base pair compound is DNA. A probe or primer length of at least 15 base pairs is dictated by theoretical and practical considerations. *See e.g.* B. Wallace and G. Miyada, "Oligonucleotide Probes for the Screening of Recombinant DNA Libraries," In Methods in Enzymology, Vol. 152, 432-442, Academic Press (1987).

The probes and primers of this invention can be prepared by methods well known to those skilled in the art (*See e.g.* Sambrook *et al. supra*). In a most preferred embodiment these probes and primers are synthesized by the polymerase chain reaction (PCR).

The present invention also relates to recombinant DNA cloning vectors and expression vectors comprising the nucleic acids of the present invention. Preferred nucleic acid vectors are those which comprise DNA. The skilled artisan understands that choosing the most appropriate cloning vector or expression vector depends on a number of factors including the availability of restriction enzyme sites, the type of host cell into which the vector is to be transfected or transformed, the purpose of the transfection or transformation (*e.g.*, stable transformation as an extrachromosomal element, or integration into the host chromosome), the presence or absence of readily assayable or selectable markers (*e.g.*, antibiotic resistance and metabolic markers of one type and another), and the number of gene copies desired in the host cell.

Vectors suitable to carry the nucleic acids of the present invention comprise RNA viruses, DNA viruses, lytic bacteriophages, lysogenic bacteriophages, stable bacteriophages, plasmids, viroids, and the like. The most preferred vectors are plasmids.

Host cells harboring the nucleic acids disclosed herein are also provided by the present invention. A preferred host is *E. coli* which has been transfected or transformed with a vector that comprises a nucleic acid of the present invention.

The present invention also provides a method for constructing a recombinant host cell capable of expressing an ORF disclosed herein, said method comprising transforming or otherwise introducing into a host cell a recombinant DNA vector that comprises an isolated DNA sequence which encodes said ORF. The preferred host cell is any strain of *E. coli* which can accomodate high level expression of an exogenously introduced gene. Transformed host cells are cultured under conditions well known to skilled artisans such that said ORF

is expressed, thereby producing the encoded protein in the recombinant host cell.

For the purpose of discovering new inhibitors of cell wall biosynthesis, it would be desirable to determine agents that inhibit enzymes required for synthesis of the cell wall and/or agents that interact with membrane proteins. A method for identifying compounds that interact with such enzymes and membrane proteins comprises contacting said proteins with a test compound and monitoring an interaction and/or inhibition by any suitable means.

The instant invention provides a screening system for compounds that interact with membrane proteins of this invention, said screening system comprising the steps of:

a) preparing a membrane protein, or membranes enriched in said protein;

b) exposing the protein source of (a) to a test compound; and

c) quantifying the interaction of said protein with said compound by any suitable means.

The screening method of this invention may be adapted to automated procedures such as a PANDEX® (Baxter-Dade Diagnostics) system, allowing for efficient high-volume screening of compounds.

In a typical screening protocol, a protein to be tested is prepared as described herein, preferably using recombinant DNA technology. A test compound is introduced into a reaction vessel containing said protein. The reaction/interaction of said protein and said compound is monitored by any suitable means. For example, a radioactively-labeled or chemically-labeled compound or

protein may be used. Specific association between a test compound and protein is monitored by any suitable means.

The following examples more fully describe the present invention. Those skilled in the art will recognize that the particular reagents, equipment, and procedures described are merely illustrative and are not intended to limit the present invention in any manner.

EXAMPLE 1

Vector for Expressing *S. pneumoniae* ORF in a Host Cell

An expression vector suitable for expressing a *S. pneumoniae* gene or fragment thereof in a variety of procaryotic host cells, such as *E. coli*, is easily made. A suitable parent vector contains an origin of replication (Ori), a marker for selecting transformants, for example, an ampicillin resistance gene (Amp), and further comprises suitable transcriptional and translational signals, for example, the T7 promoter and T7 terminator sequences, in operable-linkage to a *S. pneumoniae* coding region. For example, pET11A (obtained from Novogen, Madison WI) is linearized by restriction with endonucleases NdeI and BamHI. Linearized pET11A is ligated to a DNA fragment bearing NdeI and BamHI sticky ends and comprising a coding region for a *S. pneumoniae* ORF.

The ORF used in this construction may be modified at the 5' end (amino terminus of encoded protein or peptide) to simplify purification of the encoded protein or peptide. For this purpose, an oligonucleotide encoding 8 histidine residues is inserted after the transcriptional and translational start sites. Placement of the histidine residues at the amino terminus of the encoded protein enables the IMAC one-step protein purification procedure.

Example2

Recombinant Expression and Purification of a Protein Encoded by a *S. pneumoniae* ORF

An expression vector that carries an ORF from the *S. pneumoniae* genome, as disclosed in Example 1, and which ORF is operably-linked to an expression promoter, is transformed into *E. coli* BL21 (DE3)(*hsd*S *gal* lcIts857 *ind*1*Sam*7*nin*5*lac*UV5-T7gene 1) using standard methods. Transformants, selected for resistance to ampicillin, are chosen at random and tested for the presence of the vector by agarose gel electrophoresis using quick plasmid preparations. Colonies that contain the vector are grown in L broth and the protein produced by the vector-borne ORF is purified by IMAC, essentially as described in US Patent 4,569,794.

Briefly, the IMAC column is prepared as follows. A metal-free chelating resin (e.g. Sepharose 6B IDA, Pharmacia) is washed in distilled water to remove preservatives and then infused with a suitable metal ion [*e.g.* Ni(II), Co(II), or Cu(II)] by adding a 50mM metal chloride or metal sulfate aqueous solution until about 75% of the interstitial spaces of the resin are saturated with colored metal ion. The column is then ready to receive a crude cellular extract containing the recombinant protein product.

Unbound proteins and other materials are removed by washing the column with any suitable buffer, pH 7.5. Bound protein is eluted in any suitable buffer at pH 4.3, or preferably with an imidizole-containing buffer at pH 7.5.

Example 3

DNA Chip Production

Any one or more of the *S. pneumoniae* genome DNA fragments disclosed herein, or fragments thereof,are arrayed onto a solid support. It is preferred that fragments be in

the size range of 14 base pairs to 500 base pairs. The DNA samples are most conveniently synthesized by PCR using standard methods to amplify regions disclosed by the genomic sequences herein. The method of Schena *et al.* is used to spot about 1 ng to 10 ng of a DNA sample onto glass microscope slides that have been treated with poly-L-lysine (M. Schena *et al.* "Quantitative monitoring of gene expression patterns with a complementary DNA microarray" Science, 270, 467-470, 1995). After spotting DNA samples onto the chip and air-drying, the chips are rehydrated by incubation for about 2 hours in a humid chamber. Chips are then placed at 100° C for 1 minute, rinsed in 0.1% SDS, and treated with 0.05% succinic anhydride in 50% 1-methyl-2-pyrrolidinone and 50% boric acid.

Example 4

S. pneumoniae Gene Expression Analysis using DNA Chips

RNA prepared from cells grown under any desireable conditions is used to prime cDNA synthesis by reverse transcription, using methods well known to the skilled artisan (*See e.g.* Molecular Cloning, 2d Ed. J.Sambrook, E. Fritsch, T. Maniatis, 1989). For example, total RNA of strain R6 is prepared according to the method of Logeman *et.al.*, (*Analytical Biochemistry*, 1987, 163, 16-20) using guanidine hydrochloride. After ethanol precipitation, the total RNA is dissolved in a buffered solution such as Tris-EDTA (TE). Complementary DNA's are synthesized with the aid of the StrataScript RT-PCR kit (Stratagene,Inc.) in accordance with the supplier's recommendations (*See* Schena *et al. Id.*). Briefly, a 50 ul reaction contains about 0.1 ug/ul of RNA. First strand synthesis is primed using random primers, 1X first strand buffer, 0.03 U/ul ribonuclease block, 500 uM dATP, 500 uM dTTP, 40 uM dGTP, 40 uM fluorescein-12-dCTP (New England Nuclear), and 0.03 U/ul

reverse transcriptase. Reactions are incubated for 60 minutes at 37° C, precipitated with ethanol, and resuspended in 10 ul TE pH 8. Samples are heated for 3 minutes at 94° C and chilled on ice. The RNA is degraded by adding 0.25 ul of 10 N NaOH, followed by a 10 minute incubation at 37° C. The samples are neutralized with 2.5 ul of 1M Tris-HCl, pH 8 and 0.25 ul of 10 N HCl. After ethanol precipitation, the nucleic acid pellet is washed and dried *in vaccuo*.

Prior to hyrbrization, DNA chips prepared as in Example 3 are denatured by heating to 90°C for 2 minutes. Hybridization reactions contain about 1 ul of fluorescently-labeled cDNA, and 1 ul of hybridization buffer (10x SSC and 0.2% SDS). Probe mixtures are transferred to the surface of the chip, covered with a cover slip, and incubated for 18 hours at 65° C. Chips are washed 5 minutes at room temperature in 1X SSC, 0.1% SDS, then for 10 minutes at room temperature in 0.1X SSC, 0.1% SDS. After hybridization, chips are scanned with a laser-scanning device.

Example 5

A DNA Bio Chip for mutation analysis

Duplicate DNA chips are prepared as in Example 3. Each chip is overlayed with *S. pneumoniae* cells in a semi-solid medium, wherein said cells carry a temperature-sensitive (ts) mutation in a gene required for autolytic activity (Lyt^-). This mutation leads to resistance to lysis at 37° C, but sensitivity to lytic treatments at 30° C.

S. pneumoniae strain cwl is resistant to lysis by detergent and penicillin when grown at 37° C, but remains sensitive when grown at 30° C (cwl is derived from strain R6; *See* P. Garcia *et al.* "Mutants of Streptococcus pneumoniae that contain a temperature-senstive autolysin" *J. Gen. Microbiol.* 132, 1401-05, 1986). Strain cwl is grown at 30° C and competent cells are prepared according to any suitable method (*e.g.* LeBlanc *et.al. Plasmid* 28, 130-145, 1992; Pozzi *et al. J. Bacteriol.*178, 6087-6090, 1996). Competent cwl cells are harvested by centrifugation and resuspended at about 10^5 cells per ml in 1% melted agar supplemented with 0.1% (w/v) yeast extract (Difco) and containing 1% to 2% Triton X-100. Approximately 100 ul to 500 ul of the cell mixture is deposited per square centimeter onto the bio chip by pipetting onto the chip surface. After solidification of the agar layer, one of the bio-chips is incubated at 37° C and the other at 30° C. Cells that take up a complementing genomic DNA fragment from the chip surface will be lysed at both 30° C and 37° C, while non-complemented cells are lysed only at 30° C. Cells that are complemented by the bio-chip are recognizable by this phenotypic difference and can be further purified by well known methods.

CLAIMS

1. An isolated nucleic acid compound comprising a sequence identical to or substantially identical to a sequence selected from the group consisting of SEQ ID NO:1 through SEQ ID NO:86.

2. An isolated nucleic acid compound comprising a sequence identical to or substantially identical to a sequence selected from the group consisting of SEQ ID NO:87, SEQ ID NO:89, SEQ ID NO:91, SEQ ID NO:93, SEQ ID NO:95, SEQ ID NO:97, SEQ ID NO:99, SEQ ID NO:101, SEQ ID NO:103, SEQ ID NO:105, SEQ ID NO:107, SEQ ID NO:109, SEQ ID NO:111, SEQ ID NO:113, SEQ ID NO:115, SEQ ID NO:117, SEQ ID NO:119, and SEQ ID NO:121.

3. A substantially purified protein or fragment thereof from *S. pneumoniae* wherein said protein is selected from the group consisting of SEQ ID NO:88, SEQ ID NO:90, SEQ ID NO:92, SEQ ID NO:94, SEQ ID NO:96, SEQ ID NO:98, SEQ ID NO:100, SEQ ID NO:102, SEQ ID NO:104, SEQ ID NO:106, SEQ ID NO:108, SEQ ID NO:110, SEQ ID NO:112, SEQ ID NO:114, SEQ ID NO:116, SEQ ID NO:118, SEQ ID NO:120, SEQ ID NO:122, and SEQ ID NO:123 through SEQ ID NO:228.

4. An isolated nucleic acid compound encoding any one of the proteins or fragments thereof of Claim 3.

5. A vector comprising any one of the nucleic acid compounds of claims 1, 2, or 4.

6. A recombinant host containing any one of the vectors of claim 5.

7. A substantially purified protein from *Streptococcus pneumoniae* as in Claim 3 wherein said protein is an external target protein selected from Table 1.

8. A substantially purified protein from *Streptococcus pneumoniae* as in Claim 3 wherein said protein is a hypothetical protein selected from Table 1.

9. A substantially purified protein from *Streptococcus pneumoniae* as in Claim 3 wherein said protein is a cell wall synthetic protein selected from Table 1.

10. A substantially purified protein from *Streptococcus pneumoniae* as in Claim 3 wherein said protein is a minimal gene set protein selected from Table 1.

11. A DNA chip having arrayed thereon any at least 15 base pair fragment of any one or more of the nucleic acids of claim 1.

12. A DNA chip having arrayed thereon any at least 15 base pair fragment of any one or more of the nucleic acids of claim 2.

13. A method for evaluating gene expression in *Streptococcus pneumoniae* comprising the step of incubating a DNA chip of claim 11 or Claim 12 with cDNA prepared from *Streptococcus pneumoniae* under conditions suitable for hybridization of complementary nucleic acid sequences.

14. A computer readable medium having recorded thereon any one or more of the nucleotide sequences of Claims 1 or Claim 2.

15. A method for identifying virulence genes in *S. pneumoniae*, comprising the steps of:

a) preparing a DNA chip as in claim 11,

b) preparing labeled cDNAs from

i) *S. pneumoniae* cells recovered from an in vivo environment, and

ii) *S. pneumoniae* cells grown in vitro,

c) hybridizing individually the cDNAs of steps (b)(i) and (b)(ii) to a chip of step (a); and

d) identifying a genomic DNA fragment or fragments on said chip that hybridize to the cDNAs of (b)(i) but not with the cDNAs of (b)(ii).

16. An antibody that selectively binds to a protein or peptide of Claim 3.

17. An antibody that selectively binds to an external target protein, or fragment thereof, identified in Table 1.

18. A DNA chip of Claim 11 or Claim 12 further comprising a layer of *S. pneumoniae* cells wherein said layer contacts with said nucleic acids.

INTERNATIONAL SEARCH REPORT	International application No. PCT/US97/22578

A. CLASSIFICATION OF SUBJECT MATTER

IPC(6) :Please See Extra Sheet.
US CL :536/23.7, 23.1; 530/350, 387.1; 435/320.1, 252.3, 6; 360/1

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

U.S. : 536/23.7, 23.1; 530/350, 387.1; 435/320.1, 252.3, 6; 360/1

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

STN, APS, chip#, dna#, oligo?
MPSRCH of nucleic acid and amino acid sequence databases of IntelliGentics (SEQ ID NOs 1-228)

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	FUKUNAGA et al. Expression cloning of a receptor for murine granulocyte colony-stimulating factor. Cell. 20 April 1990, Vol. 61, pages 341-350, see entire document.	11-13 and 15
Y	GEETHA-HABIB et al. Glycosylation site binding protein, a component of oligosaccharyl transferase, is highly similar to three other 57 kd liminal proteins of the ER. Cell. 23 September 1988, Vol. 54, pages 1053-1060, see entire document.	11-13 and 15
Y	STRAUSS et al. Complete nucleotide sequence of the genomic RNA of Sindbis virus. Virology. February 1984, Vol. 133, No. 1, pages 92-110, see entire document.	11-13 and 15

[X] Further documents are listed in the continuation of Box C. [ ] See patent family annex.

* Special categories of cited documents:

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier document published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
25 MARCH 1998	09 APR 1998
Name and mailing address of the ISA/US Commissioner of Patents and Trademarks Box PCT Washington, D.C. 20231 Facsimile No. (703) 305-3230	Authorized officer JAMES MARTINELL Telephone No. (703) 308-0196

Form PCT/ISA/210 (second sheet)(July 1992)★

C (Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	WILSON et al. 2.2 Mb of contiguous nucleotide sequence from chromosome III of C. elegans. Nature. 03 March 1994, Vol. 368, No. 6466, pages 32-38, see entire document.	11-13 and 15
Y	RESTREPO et al. Antigen diversity in the bacterium B. hermsii through "somatic" mutations in rearranged vmp genes. Cell. 09 September 1994, Vol. 78, pages 867-876, see entire document.	11-13 and 15
Y	WHITEWAY et al. Dominant negative selection of heterologous genes: Isolation of Candida albicans genes that interfere with Saccharomyces cerevisiae mating factor-induced cell cycle arrest. Proc. Natl. Acad. Sci. USA. 15 October 1992, Vol 89, No. 20, pages 9410-9414, see entire document.	11-13 and 15
Y	LACKS et al. Genetic basis of the complementary DpnI and DpnII restriction systems of S. pneumoniae: An intracellular cassette mechanism. Cell. 26 September 1986, Vol. 46, pages 993-1000, see entire document.	11-13 and 15
Y	STEENBERGH et al. Structure and expression of the human calcitonin/CGRP genes. FEBS Letters. 01 December 1986, Vol. 209, No. 1, pages 97-103, see entire document.	11-13 and 15
Y	SHAMBAUGH et al. The spliceosomal U small nuclear RNAs of Ascaris lumbricoides. Molecular and Biochemical Parasitology. April 1994, Vol. 64, No. 2, pages 349-352, see entire document.	11-13 and 15
Y	MAXWELL et al. A novel NF-kappaB p65 spliced transcript lacking exons 6 and 7 in a non-small cell lung carcinoma cell line. Gene. 12 December 1995, Vol. 166, No. 2, pages 339-340, see entire document.	11-13 and 15
Y	KONYECSNI et al. Broad-host-range plasmid and M13 bacteriophage-derived vectors for promoter analysis in Escherichia coli and Pseudomonas aeruginosa. Gene. 30 December 1988, Vol. 74, No. 2, pages 375-386, see entire document.	11-13 and 15
Y	ALLIKMETS et al. Cloning and organization of the abc and mdl genes of Escherichia coli; Relationship to eukaryotic multidurg resistance. Gene. 22 December 1993, Vol. 136, Nos. 1 and 2, pages 231-236, see entire document.	11-13 and 15
Y	SCHWEIZER. Two plasmids, X1918 and Z1918, for easy recovery of the xylE and lacZ reproter genes. Gene. 30 November 1993, Vol. 134, No. 1, pages 89-91, see entire document.	11-13 and 15

INTERNATIONAL SEARCH REPORT

International application No. PCT/US97/22578

C (Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	KRAFT et al. Sequence of the complete P protein gene and part of the M protein gene form the histidine transport operon of Escherichia coli compared to that of Salmonella typhimurium. Nucleic Acids Res. 26 October 1987, Vol. 15, No. 20, page 8568, see entire document.	11-13 and 15
Y	SCHULER et al., Characterization of the human gene encoding LBR, an integral protein of the nuclear envelope inner membrane., J. Biol. Chem. 15 April 1994, Vol. 269, No. 15, pages 11312-11317, see entire document.	11-13 and 15
Y	LOTTENBERG et al. Cloning, sequence analysis, and expression in Escherichia coli of a streptococcal plasmin receptor. J. Bacteriol. August 1992, Vol. 174, No. 16, pages 5204-5210, see entire document.	11-13 and 15
Y	COPPOLA et al. Sequence and transcriptional activity of the Escherichia coli K-12 chromosome region between rrnC and ilvGMEDA. Gene. 02 January 1991, Vol. 97, No. 1, pages 21-27, see entire document.	11-13 and 15
Y	MUTO et al. Preferential use of A- and U-rich codons for Mycoplasma capricolum ribosomal proteins S8 and L6. Nucleic Acids Res. 1984, Vol. 12, No. 12, pages 8209-8217, see entire document.	11-13 and 15
Y	RUDNER et al. The spo0K locus of Bacillus subtilis is homologous to the oligopeptide permease locus is required for sporulation and competence. J. Bacteriology. February 1991, Vol. 173, No. 3, pages 1388-1398, see entire document.	11-13 and 15
Y	GREEN et al. Isolation and characterisation of a bovine cDNA encoding eukaryotic initiation factor 2α. Biochim. Biophys. Acta. 08 October 1991, Vol. 1090, No. 2, pages 277-280, see entire document.	11-13 and 15
Y	LOGEMANN et al. Modes of expression and common structural features of the complete phenylalanine ammonia-lyase gene family in parsley. Proc. Natl. acad. Sci. USA. 20 June 1995, Vol. 92, No. 13, pages 5905-5909, see entire document.	11-13 and 15
Y	SHERWOOD et al. Characterization of HIR1 and HIR2, two genes required f r regulation of histone gene transcription in Saccharomyces cerevisiae. Molec. Cell. Biol. January 1993, Vol. 13, N . 1, pages 28-38, see entire document.	11-13 and 15

C (Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	GAGNON et al. Cloning, sequencing and expression in Escherichia coli of the ptsI gene encoding enzyme I of the phosphoenolpyruvate:sugar phosphotransferase transport system from Streptococcus salivarius. Gene. 02 November 1992, Vol. 121, No. 1, pages 71-78, see entire document.	11-13 and 15-17
Y	PUYET et al. Characterization of the Streptococcus pneumoniae maltosaccharide regulator MalR, a member of the LacI-GalR family of repressors displaying distinctive genetic features. J. Biol. Chem. 05 December 1993, Vol. 268, No. 34, pages 25402-25408, see entire document.	11-13 and 15
Y	CROSS et al. Purification of CpG islands using a methylated DNA binding column. Nature Genetics. March 1994, Vol. 6, pages 236-244, see entire document.	11-13 and 15
Y	SALUJA et al. The genetic basis of colony opacity in Streptococcus pneumoniae: evidence for the effect of box elements on the frequency of phenotypic variation. Molecular Microbiology. April 1995, Vol. 16, No. 2, pages 215-227, see entire document.	11-13 and 15-17
Y	ROTHNIE et al. Retrotransposon-like nature of Tp1 elements: implications for the organisation of highly repetitive, hypermethylated DNA in the genome of Physarum polycephalum. Nucleic Acids Res. 25 January 1991, Vol. 19, No. 2, pages 279-286, see entire document.	11-13 and 15
Y	GITT et al, A strong sequence homology exists between the major RNA polymerase δ factors of Bacillus subtilis and Escherichia coli. J. Biol. Chem. 25 June 1985, Vol. 260, No. 12, pages 7178-7185, see entire document.	11-13 and 15
Y	LEUVEN et al. Structure of the gene (LRP1) coding for the human α2-macroglobulin receptor lipoprotein receptor-related protein. Genomics. 01 November 1994, Vol. 24, No. 1, pages 78-89, see entire document.	11-13 and 15
Y	YOSHIOKA et al. Nucleotide sequence of the promoter-distal region of the tra operon of plasmid R100, including traI (DNA helicase I) and traD genes. J. Mol. Biol. 05 July 1990, Vol. 214, No. 1, pages 39-53, see entire document.	11-13 and 15
Y	HUI et al. Genetic transformation in Streptococcus pneumoniae: Nucleotide sequence analysis shows comA, a gene required for competence induction, to be a member of the bacterial ATP-dependent transport protein family. J. Bacteriol. January 1991, V 1 173, No. 1, pages 372-381, see entire document.	11-13 and 15

C (Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	BARON et al. Co-regulation of two gene activities by tetracycline via a bidirectional promoter. Nucleic Acids Res. 11 September 1995, Vol. 23, No. 17, pages 3605-3606, see entire document.	11-13 and 15
Y	BORK et al. Exploring the Mycoplasma capricolum genome: a minimal cell reveals its physiology. Molecular Microbiology. June 1995, Vol. 16, No. 5, pages 955-967, see entire document.	11-13 and 15
Y	CHAMBERS et al. The pMTL nic$^-$ cloning vectors. I. Improved pUC polylinker regions to facilitate the use of sonicated DNA for nucleotide sequencing. Gene. 15 August 1988, Vol. 68, No. 1, pages 139-149, see entire document.	11-13 and 15
Y	FENOLL et al. Molecular basis of the optochin-sensitive phenotype of pneumococcus: characterization of the genes encoding the F_0 complex of the Streptococcus pneumoniae and Streptococcus oralis H^+ -ATPases, Molecular Microbiology, May 1994, Vol. 12, No. 4, pages 587-598, see entire document.	11-13 and 15-17
Y	STEGLITZ-MORSDORF et al. Cloning, heterologous expression, and sequencing of the Proteus vulgaris glnAntrBC operon and implications of nitrogen control on heterologous urease expression. FEMS Microbiology Letters. 15 January 1993, Vol. 106, No. 2, pages 157-164, see entire document.	11-13 and 15
Y	DE WIT et al. Structure of the gene for the human myeloid IgA Fc receptor (CD89). J. Immunol. 01 August 1995, Vol. 155, No. 3, pages 1203-1209, see entire document.	11-13 and 15
Y	WEBSTER et al. Nucleotide sequence of the galactose gene cluster of Kluyveromyces lactis, Nucleic Acids Res. 25 August 1988, Vol. 16, No. 16, pages 8192-8194, see entire document.	11-13 and 15
Y	KIRSCH et al. Chemotactic methylesterase promotes adaptation to high concentrations of attractant in Bacillus subtilis. J. Biol. Chem. 05 September 1993, Vol. 268, No. 25, pages 18610-18616, see entire document.	11-13 and 15
Y	YAMASHITA et al. Molecular characterization of a Streptococcus mutans mutant altered in environmental stress responses. J. Bacteriol. October 1993, Vol. 175, No. 19, pages 6220-6228, see entire document.	11-13 and 15-17
Y	FRASER et al. The minimal gene complement of Mycoplasma genitalium. Science. 20 October 1995, Vol. 270, pages 397-403, see entire document.	11-13 and 15

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US97/22578

C (Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	TOMINAGA et al. Upstream region of a genomic gene for human mitochondrial transcription factor 1. Biochim. Biophys. Acta. 15 June 1992, Vol. 1131, No. 2, pages 217-219, see entire document.	11-13 and 15
Y	CORTI et al. Molecular cloning of cDNAs encoding human carnitine acetyltransferase and mapping of the corresponding gene to chromosome 9q34.1. Genomics. 01 September 1994, Vol. 23, No. 1, pages 94-99, see entire document.	11-13 and 15
Y	LUSIS et al. Cloning and expression of apolipoprotein B, the major protein of low and very low density lipoproteins. Proc. Natl. Acad. Sci. USA. July 1985, Vol. 82, No. 14, pages 4597-4601, see entire document.	11-13 and 15
Y	KOHLBRECHER et al. Staphylococcal phosphoenolpyruvate-dependent phosphotransferase system: Molecular cloning and nucleotide sequence of the Staphylococcus carnosus ptsI gene and expression and complementation studies of the gene product. J. Bacteriol. April 1992, Vol. 174, No. 7, pages 2208-2214, see entire document.	16 and 17
Y	FLEISCHMANN et al. Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. Science. 28 July 1995, Vol. 269, pages 496-512, see entire document.	11-13 and 15-17
Y	PEASE et al. Light-generated oligonucleotide arrays for rapid DNA sequence analysis. Proc. Natl. Acad. Sci. USA. 24 May 1994, Vol. 91, No. 11, pages 5022-5026, see entire document.	11-13 and 15
X,P	CHENG et al. The com locus controls genetic transformation in Streptococcus pneumoniae. Molecular Microbiology. February 1997, Vol. 23, No. 4, pages 683-692, see entire document.	1, 4-6, and 14
Y	MORIARTY et al. Antibodies to peptides detect new hepatitis B antigen: Serological correlation with hepatocellular carcinoma. Science. 25 January 1985, Vol. 227, pages 429-433, see entire document.	16 and 17
X	WO 96/10647 A (FUSO PHARMACEUTICAL INDUSTRIES, LTD.) 11 April 1996, see entire document.	1, 4-6, and 14
X --- Y	ALLOING et al. The ami locus of the Gram-positive bacterium Streptococcus pneumoniae is similar to binding protein-dependent transport operons of Gram-negative bacteria. Molecular Microbiology. 1990, Vol. 4, No. 4, pages 633-644, see entire document.	1, 4-6, and 14 ---------- 16 and 17

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US97/22578

A. CLASSIFICATION OF SUBJECT MATTER:
IPC (6):

C12N 15/31, 15/00, 1511; C07K 14/315; C07H 21/04; A61K 39/40, 39/395; G01N 33/569; G11B 5/00